Computing Technologies and Methods for NEEScomm

Ruth Pordes, October 2014

Abstract: This report gives some technologies and methods in place today or emerging in scientific and research cyber-infrastructures, computing and software projects, in the US and Europe, that may be relevant to the future of NEES computing.

Acknowledgements: Useful discussions and input are appreciated from: Andre Barbosa Oregon State University; Dan Stanzione, Texas Advanced Supercomputing Center; Tim Ahern, Incorporated Research Institutions for Seismology; John Cobb, Oakridge National Laboratory; Shirley Dyke and Thomas Hacker of NEEScomm; Phil Maechlin, SCEC; David Lifka Cornell University Center for Advanced Computing, and Michael McLennan of Hubzero.

Table of Contents

1 Introduction

NEES computing provides mature capabilities for: Diverse data ingestion, curation and classification; Rich search for, access to and manipulation of data across the stored datasets; and access to resources from simple to extremely complex computations. This

infrastructure is based around the HUBzero[1] software system. The NEES community already makes significant use of the following tools: FlexTPS[2], OpenSees[3], OpenFresco[4], RDV[5]. Development of HUBzero is ongoing. In fall 2014 support for Google Drive, Incommon authentication, Redcloud, and Galaxy are integrated into the portal and either being used by early adopters or in test.

The purpose of this report is to provide pointers to additional technology and methods that could be integrated into, or used as a pattern to extend, the NEES computing system to the benefit of NEES researchers and communities.

2 Data Movement and Storage

One main focus of NEEShub is the transfer and storage of streaming and file based data/information. Easy to use and robust data file and streaming data movement are of importance. A few general items of note:

- Network aware data movement is at the forefront of research.
- Managing the "last 100 meters" "host interfaces" and monitoring. Perfsonar is a joint project between Internet 2, ESNET and GEANT in Europe that supplies monitoring software and guidance on hardware at the site border on which to install it.[6]
- Climate scientists have a useful evaluation of existing and potential data movement technologies that includes some of those below and recommendations for GridFTP and BBCP[7] below.
- Distributed file systems continue to be the "nirvana" of access to data over the wide area, but to date have not given the needed throughput and robustness for the scientific community.

2.1 Commodity and Commercial Solutions

2.1.1 ASPERA

While the commercial space is not yet accommodating the scalability of the scientific domain new offerings are starting to appear. Aspera's faspex is being tested for use across cloud interfaces that provide an Amazon AWS interface. http://cloud.asperasoft.com/fileadmin/user_upload/documents/Aspera_faspex_Datasheet_01.pdf recommended for use across heterogeneous cloud systems AWS.

[1] https://HUBzero.org/
[2] https://nees.org/resources/flextps
[3] http://opensees.berkeley.edu/
[4] http://openfresco.berkeley.edu/
[5] https://nees.org/resources/rdv
[6] http://www.es.net/about/esnet-staff/advanced-network-technologies/Brian-Tierney/
[7] http://www.copernicus.eu/pages-principales/library/presentations/copernicus-big-data-workshop/international-perspectives/?no_cache=1&cHash=4cae5bf1706351cb2839cf1e3306e12f

2.1.2 DropBox

http://www.dropbox.com/

Dropbox is a very popular service to share and store files of increasing size. The ease of installation, the number of platforms supported, and the accessibility from the web, mobile and other platforms makes it a popular choice for file storage and access of medium size. Its limitations come with the lack of APIs for integration into scientific applications, the bandwidth limitations for data transfer and the ability to configure security and access to the user communities needs.

2.1.3 Google Drive

https://drive.google.com/

Google Drive offers similar functionality to DropBox. It is not as mature or proven to be as scalable at this time.

2.2 Open Source Community Solutions

2.2.1 Bittorrent

http://en.wikipedia.org/wiki/BitTorrent, While not initially designed for scientific data developments to improve it usability and performance for LHC sized data continue. Evaluations and tests are being done as part of the R&D for future data taking runs[8] to investigate if real-time requests accompanied by bulk streaming data transfer can meet throughput needs of analysis programs running remotely from the large data stores.

2.2.2 Bulk Backup Copy Program (BBCP)

http://www.slac.stanford.edu/~abh/bbcp/

bbcp is a multi-streaming point-to-point network file copy application with excellent network transfer rates. It is supported by some of the HPC systems and has been shown to be high throughput and "simple" for point to point needs[9]. The application was originally written for transferring large files of the data-intensive High-Energy Physics community. Unlike other utilities available, bbcp only requires that an sshd server be setup for a machine. As most (if not all) machines will already be configured to allow people to ssh to them there is no additional setup required from a systems administrator. You simple need to obtain a copy of bbcp and install it using your account

2.2.3 Globus Online

https://www.globus.org/file-transfer

Globus has been adopted by the US NSF and DOE high performance computing systems.

---

[8] https://cds.cern.ch/record/1630387/files/s0129183114300012.pdf

https://indico.fnal.gov/conferenceDisplay.py?confId=8389

[9] https://www.olcf.ornl.gov/kb_articles/transferring-data-with-bbcp/

It provides high-speed configurable file transfer service between any two end-points, optionally managed by a third end-point; with a diversity of authentication protocols supported. For high throughput sustained use there is a cost model that is referred to below. There is ongoing work to ensure the implementation of the file transfer layer allows use of RDMA and other modern protocols. Globus recently supports data publishing and discovery features.

2.2.4 National Data Service

http://www.nationaldataservice.org/

This is a new consortium that aims to support scientists and researchers across all disciplines to find, reuse, and publish data. Current technology components, to be used in the Materials Data Facility[10], the first repository to be supported as part of the consortium, include storage environments at NCSA and at Argonne National Laboratory and data management from Globus.

2.2.5 Open Source DataTurbine (OSDT) Initiative

http://www.dataturbine.org

The OSDT software extensions the basic functionality of the Data Turbine streaming data capabilities to include sensor probes, monitoring capabilities, cloud interfaces and other functions. This system is being considered by Earthcube, as well as NEEScomm, to support data streaming.

2.2.6 UDP-base Data Transfer (UDT)

http://udt.sourceforge.net/news.html

UDT is a UDP based data transfer program for wide area networks. Because it is based on UDP additional software is needed to provide the guaranteed delivery. But also the transfer rates are improved over those for TCP/IP and bandwidth challenge successes at SC08 and SC09 were based on UDT. UDT is one of the functions offered by the Open Science Data Cloud consortium. The source repository has been quiet for the past few years.

2.3 Distributed Shared File Systems

2.3.1 General Parallel File System (GPFS)

http://www-03.ibm.com/systems/platformcomputing/products/gpfs/

GPFS is a commercially available IBM product that provides a high-performance clustered file system, again with wide area capabilities that are not fully matured.

2.3.2 Global Federated File System (GFFS)

http://genesis2.virginia.edu/wiki/Main/GFFS

[10] https://wiki.ncsa.illinois.edu/display/NDS/Pilot%3A+Materials+Data+Facility

GFFS is being developed and deployed by the NSF XSEDE project Genesis II. GFFS provides access to and manipulation of remote file systems by employing a global path-based namespace. Transparent access to data is realized by using OS-specific file system drivers.

2.3.3 Lustre

www.lustre.org

Lustre is well supported as an institutional-wide file system and work is ongoing to provide equivalent capabilities over the wide area network. When large numbers of files – especially small files - are being transferred or accessed the performance is reduced.

3 Analytics and Portals

In this section we touch on projects that are developing and supporting generalized data analytic tools and methods, and web based digital portals for the scientific research community. We do not include HUBzero as NEES is already based on this solution.

3.1 Agency Programs

3.1.1 DOE SCiDAC Institutes

http://www.scidac.gov/institutes.html

The DOE SCiDAC institutes research and develop algorithms, methods, and scientific software tools to advance scientific discovery through modeling and simulation. Of particular note is:

Scalable Data Management, Analysis, and Visualization (SDAV, http://sdav-scidac.org/) provides a suite of data analytics and visualization tools, and works with the application communities to integrate and extend them. The toolkit is being supported on a variety of modern compute hardware. Technologies of potential interest include an advanced data visualization tool (VISIT) for analyzing very large datasets, and algorithms (FASTBIT) for quickly finding records satisfying user-specified conditions from large, complex data sets.

3.1.2 NNSA Predictive Science Academic Alliance Program (PSAAP)

http://nnsa.energy.gov/aboutus/ourprograms/defenseprograms/futurescienceandtechnologyprograms/asc/univpartnerships/psaap

The five PSAAP centers aim is to research, develop, and establish tools and methods to support validated, large-scale, multidisciplinary, simulation-based “Predictive Science” capabilities in Universities and Laboratories. In particular, the UQ Toolkit[11] methods are a well-regarded collection of libraries and tools for the quantification of uncertainty in numerical model predictions.

[11] http://www.sandia.gov/UQToolkit/

3.1.3 NSF Scientific Gateways Institute

http://sciencegateways.org/

This is an NSF funded planning (conceptualization) forum to develop the understanding and scope of a potential Institute. The institute would provide coordination of and a central point for consulting, support and requirements gathering for scientific communities needing to provide and use application web portals for data and computation.

3.2 Domain Specific Projects

3.2.1 Earthcube Building Blocks

http://workspace.earthcube.org/building-blocks

The Earthcube Building Blocks projects will provide implementations for specific cyberinfrastructure components to allow for data sharing across the geosciences:

- Deploying Web Services Across Multiple Geoscience Domains.
- Specifying and Implementing ODSIP, a Data-service Invocation Protocol.
- Software Stewardship for the Geosciences.
- A Broker Framework for Next Generation Geoscience (BCube).
- Integrating Discrete and Continuous Data.
- Leveraging Semantics and Linked Data for Geoscience Data Sharing and Discovery (OceanLink).
- A Cognitive Computer Infrastructure for Geoscience.
- Earth System Bridge: Spanning Scientific Communities with Interoperable Modeling Frameworks.
- Community Inventory of EarthCube Resources for Geoscience Interoperability (CINERGI).

3.2.2 IPLANT Collaborative

http://www.iplantcollaborative.org/

http://agaveapi.co/

iPlant provides computing and software tools and services including analysis tools and workflows, visualization and image analysis. The data model definitions and support methodologies provide a pattern that can be used by other domains and projects.

Part of IPLANT is the Agave software and API “science-as-a-service” provides a good example of an integrated toolkit,

3.3 Incorporated Research Institutions for Seismology (IRIS)

http://www.iris.edu/hq/

IRIS offers software, tools and services for scientists undertaking research including data

analysis tools (e.g. time series processing), data classification and metadata, and web-based access to community data.

3.3.1 Ocean Observatories Initiative Cyberinfrastructure (OOI)

http://ci.oceanobservatories.org/

The cyberinfrastructure for OOI is called the Integrated Observatory Network (ION). It includes hardware and software to set up observatories, platforms, instruments, as well as the fundamental data processing and data product services that provide scientists the ocean measurements and calculations. The primary language in use is Python and communications are through AMQP messaging[12] system.

3.3.2 Southern California Earthquake Center (SCEC)

http://www.scec.org/

SCEC develops and sustains a variety of general software for the analysis and display of earthquake simulation, modeling and prediction information. Of note is OpenSHA, an open-source platform for conducting and Seismic Hazard Analysis, and the SCEC Virtual Display of Objects (SCEC-VDO) data visualization and exploration toolkit.

3.3.3 Sustainable Environment for Actionable Data

http://sead-data.net/

This is an NSF funded project to create data services designed to meet the needs of sustainability science research. The tools, which are in development and not yet released for general use, include: Harvesting of ORCID (see below) information; and tools for the provision, management and intelligent access to a virtual archive of environmental data.

3.3.4 Systems Biology Knowledgebase (KBASE)

http://www.kbase.us/

The DOE sponsored KBase provides a computational framework and tools for integrating and analyzing large, diverse datasets generated by the biological scientific community. It has a service oriented architecture, design and functional components that can provide a pattern for other similar projects.

3.4 General Open Source Toolkits

3.4.1 HYDRA

http://projecthydra.org/

[12] http://www.amqp.org/

Hydra is a community-based project that includes collaboration between Fedora Commons ad Duraspace – recognize repository and repository software open source providers. The HYDRA software provides a layered suite for managing and persisting digital objects, indexing and accessing them, and developing solutions for acting on, displaying and integrating these objects into diverse applications. The Hydra eco-system provides a community based collaboration for the development, support and evolution of the tools.

3.4.2 R Toolkit

http://toolkit.pbworks.com/w/page/22358273/RToolkit

http://www.rstudio.com/

R is a free statistical software package providing free cutting-edge statistical tools for data analysis and visualization. It provides as easy to use, ubiquitously available, analysis toolkit for faculty and students and, because of its widespread use, has useful add-ons and extensions available.

RStudio gives integrated development, GUI and other packages for use with R applications.

3.4.3 ROOT Data Analysis Frameworks

http://root.cern.ch

ROOT provides a set of object-oriented frameworks with all the functionality needed to handle and analyze large amounts of data in a very efficient way. The high-energy physics community in all aspects of data analysis and presentation heavily uses it.

3.4.4 Stream Encode Explore Disseminate My Experiments (SeedMe)

http://www.seedme.org/

SeedMe provides a general command line interface and web infrastructure to share datasets and results including video, audio, and streaming data, with a maximum file size of 100 MB for any single file.

4 Scientific Telepresence and Real-Time

This is a growing field and this review does not do justice to the offerings available. We look mainly at DOE and NSF community tools and projects. The following projects are already collaborating closely with, or are indeed integrated into, the NEEScomm IT functionality: FlexTPS, Real-Time Data Viewer (RDV) and Open Source DataTurbine (OSDT) Initiative.

4.1 National Ecological Observatory Network (NEON)

http://www.neoninc.org/

The Neon software is currently being developed by a mix of in-house and externally

contracted developments. The system is to be fully deployed by 2017. Software components are called Maximo for the cyberinfrastructure and eVine for the data acquisition system. The remote sensing arm cyberinfrastructure is provided by the Airborne Observation Platform (AOP) instrumentation suite.

4.2 Android Sensor Pod and OSDT

http://www.dataturbine.org/content/osdt-android-sensorpod

Android SensorPod is a custom designed mobile computing platform for assembling wireless sensor networks that includes real time data collection and remote data collection integrated with the OSDT Data Turbine data streaming functionality. The SensorPod itself is a complete package originally intended for environmental monitoring applications but can easily be configured to fit any sensor network or system.

4.3 Experimental Physics and Industrial Control System (EPICS)

http://www.aps.anl.gov/epics/

is a set of Open Source software tools, libraries and applications developed collaboratively and used worldwide to create distributed software real-time control systems for scientific instruments such as a particle accelerators, telescopes and other large scientific experiments. It is extensively used at the Argonne Proton Synchrotron (APS), the Oakridge Spallation Neutron Source (SNS). The collaboration continues to develop extensions and additional features needed.

5 Networking

5.1 CISCO

Besides being a premier vendor of networking hardware, Cisco supports collaborative projects with scientific research groups, including networked sensor based systems. http://www.cisco.com/web/about/ac50/ac207/crc_new/index.html

5.2 Energy Sciences Network (ESnet)

http://www.es.net/

ESnet provides high-bandwidth, reliable connections between national laboratories, universities and other research institutions, funded by the DOE Office of Science. ESnet is increasingly supporting services and software to help researchers have deployed and use these connections, in particulate on demand secure circuits and reservation system (OSCARS) software and a standardized test and measurement framework (PerfSONAR). ESnet provides project specific researcher requirements and system implementation consulting and support.

5.2.1 LHCONE

ESNET has recently taken responsibility for the upgrade of the transatlantic network for the LHC Experiments in Run 2 supporting full utilization of 100G links[13]. This includes a commitment to operate networks to many university sites, as well as monitor and respond to operational issues anywhere in the global networks used for LHC data storage and analysis.

5.3 Geni Project

http://www.geni.net/

The NSF funded Geni project aims to research and then implement new network protocols for large-scale data movement. While there are research applications ongoing its deployment for large-scale production applications is still down the road.

5.4 Internet2

http://www.internet2.edu/

Internet2 provides Research and Education Networks include nonprofit organizations that are sub-state, state or multi-state in scope and have a principal mission to provide network infrastructure and services primarily to the research and education community.

5.5 The SuperComputing Conference

The Supercomputing Conference bandwidth, now Network Research, challenge[14] gives a good list of directions being pursued in advance of new production (robust, user accessible) services are available. The SC14 projects will be showcased in November in New Orleans. A representative list is shown by the list of projects at SC13:

- 100 Gbps Networks for Next Generation Petascale Science Research and Discovery
- Advanced Network Analytics at 100 Gbps
  Application-Aware Traffic Engineering for Wide Area Networks using OpenFlow
  Data-Scope at 100 Gbps Across National Data-Intensive Computational Science Testbeds
- Dynamic Monitoring and Adaptation of Data Driven Scientific Workflows Using Federated Cloud Infrastructure
- Enhancement of Globus GridFTP over SmartNIC User-Programmable 10GigE NIC
  Firewalling Science DMZ without Bottlenecks: Using Application-Aware Traffic Steering
- The Global Environment for Network Innovations (GENI) at 100 Gbps
  HPC Open Science Data Cloud (OSDC) for Data Intensive Research at 100 Gbps

---

[13] http://www.es.net/news-and-publications/esnet-news/2014/100-gbps-test-link-sets-pace-for-faster-trans-atlantic-data-transfers/

[14] http://sc13.supercomputing.org/content/scinet-network-research-exhibition

International Network Research Testbed Federation Prototypes
- Next Generation Genome Sequencing Data at 100 Gbps
- Provider Backbone Bridging Based Network Virtualization
- Using Remote I/O for Large-Scale Long Distance Data Storage and Processing
- SDN Innovation Framework Enabling Programmability of Flows within the Network

6 Cloud Computing

Use of single and federation of multiple Cloud resources across public, scientific and private is becoming a possibility giving applications the ability to move between resources as needed, to "burst" for peak usage into new types of computing. A recent survey from XSEDE[15] gives the main reasons researchers are interested in Cloud computing as "(1) on-demand access to burst resources, (2) compute and data analysis support for high throughput scientific workflows, and (3) enhanced collaboration through the rapid deployment of research team web sites and the sharing of data."

In addition to Amazon and Google commercial offerings the OpenStack community aims to form an international scientific cloud infrastructure – with Rackspace offering services in its support.

6.1 Amazon Web Services

http://aws.amazon.com/

Amazon Web Services offers a broad set of global compute, storage, database, analytics, application, and deployment services in the Cloud.

6.2 BOX

https://www.box.com/

A commercial solution offering cloud hosted storage of and global access to data. Internet2 is offering its members this service for an additional fee[16]. Many universities, including Indiana University and the University of Michigan are now including this service to their faculty and students for data in the less than 100Gigabyte range. Typically BOX is not being used for data classified as "limited access/restricted" or "critical".

6.3 Google Cloud Platform

https://cloud.google.com/

The Google Cloud Platform hosts generalized computing, storage, database and application services.

[15] https://www.xsede.org/xsede-nsf-release-cloud-survey-report
[16] http://www.internet2.edu/products-services/cloud-services-applications/box/

6.4 MicroSoft Azure

http://azure.microsoft.com/en-us/

Provides windows and Linux virtual machine resources on demand.

6.5 RackSpace Managed Cloud

http://www.rackspace.com/managed-cloud/

Rackspace offers compute and storage resources interfaced using the OpenStack open source software system.

7 High Performance Computing

NEEShub Batch interface provides an interface to and monitoring and feedback from jobs running in a distributed environment.

7.1 HPC Resources

7.1.1 DOE HPC Computing

http://www.doeleadershipcomputing.org

The DOE supports user computing facilities at NERSC, ANL and ORNL[17]. Access to the resources are made through INCITE applications.

7.1.2 Extreme Science and Engineering Discovery Environment (XSEDE)

https://www.xsede.org/

The collection of computing resources provided by XSEDE enable guaranteed allocation of compute and storage cycles and also access to opportunistically available cycles on some of the resources. New XSEDE resources, such as the SDSC Comet system coming online at the end of 2014, will provide Cloud interfaces to the resources.

7.1.3 Open Science Grid (OSG)

http://www.opensciencegrid.org

The Open Science Grid is a collaborative Consortium providing an open facility for the shared use of high throughput computing resources.

7.2 New Hardware Support

Retooling applications to be efficient on the emerging multi-core standard offerings is a major concern and effort of all scientific fields.

7.2.1 NVidia

https://research.nvidia.com/content/cuda-research-centers

[17] http://www.doeleadershipcomputing.org/incite-awards/

NVIdia provides hardware and support grants for research into the use of their hardware and software platforms.

7.2.2 Intel

http://www.intel.com/content/www/us/en/education/university/university-collaborative-research.html.html

Intel supports collaborative research opportunities with universities.

7.3 Scientific Clouds

7.3.1 The Open Science Data Cloud (OSDC)

https://www.opensciencedatacloud.org/

OSDC offers membership based computing, storage, software and support and provides cloud interfaces to all resources.

7.3.2 RedCloud Science Cloud

https://www.cac.cornell.edu/redcloud/

RedCloud offers a cloud interface to computing resources for applications that need to use the application MATLAB. It also supports use of the parallel version of the software.

7.3.3 European Grid Infrastructure

https://www.egi.eu/infrastructure/cloud/

In Europe there is a community of scientific cloud providers under the umbrella of the EGI program with Common Interfaces based on the OCCI standard.

7.4 Common Access and Support Tools

Interfacing to and adapting existing tools associated with these distributed high performance and/or cloud resources:

7.4.1 BOINC

http://boinc.berkeley.edu/

The BOINC software is being continually developed to support more computational environments. It now works on Androids and is a good demonstration of how to harvest computational cycles from Smartphones. It is wrapped in a virtualized environment and can be run on Amazon and other commercial clouds.

7.4.2 Docker

https://www.docker.com/

an open-source project that automates the deployment of applications inside software containers for use on virtualized systems.

Globus Online

for the transfer and management of files

7.4.3 XDMOD

https://xdmod.ccr.buffalo.edu/

for monitoring and accounting of jobs on XSEDE. An open source version of the software is available.

7.4.4 CIConnect

http://ci-connect.net/

CIConnect connects and interfaces to individual researcher clusters at universities.

7.4.5 Single Signon Authentication

https://www.eduroam.us/

https://www.incommon.org/

https://shibboleth.net/

Different authentication and signon environments have been acknowledged as one of the most significant barriers to and challenges for researchers to use a common cyberinfrastructure. Technologies are now well supported that allow individual Digital Identities to be used seamlessly across a diversity of computational environments. A single signon can suffice across the local university, centralized computing hubs, and distributed computational resources. The most common technologies and collaborations in the US are eduroam, Shibboleth and the InCommon federation.

7.5 Workflow Software

https://airavata.apache.org/
http://www.phylo.org/
http://galaxyproject.org/
http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/doc.prd/index.html
https://kepler-project.org/
https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA
http://pegasus.isi.edu/
http://www.taverna.org.uk/
http://guse.hu/about/architecture/ws-pgrade

There are many workflow engines, some of which are general and others are domain specific. A recent survey by Fermilab shows that Galaxy, Pegasus, Askalon, Kepler, and Taverna Triana remain popular in the research community. For HEP, Panda and GlideinWMS provide scalable underlying workload management systems. Other toolkits of interest could be Airavata portal and Cipres gateway supported by the Scientific Gateways group associated with XSEDE, and in Europe the WSPGrade portal.

8 Community Building

NEEShub already provides a mature infrastructure for building, fostering and maintaining collaborations. The use of standard groups, facebook, wikis etc as well as reference standard Document IDs assures the data and information collected are accessible and registered in the broader community. Some additional opportunities are listed below:

8.1 ORCID - researcher persistent identifiers

http://orcid.org/

"ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized."

8.2 Social Media, Bulletin Boards

Facebook, Linkedin, and other community boards are used for community building and sharing of information. A diverse set of bulletin and discussion boards are also useful. Basecamp, Google etc. are all used by some parts of the HEP Community but there is no "one size fits all" and by their very nature these are easy to use and adopt. They are not easy to adapt or integrate into other systems.

9 Standards and Organizational Practices

Participating in standards development is expensive; however benefiting from the work done and common organizational practices can help in terms of reusable software, common semantics and reminding of the scope needed. NEES already participates in fora such as the NSF Large Facilities Security Workshop, and as a science domain driving the NSF SI2 program of work.

9.1 Data preservation

A multi-disciplinary effort in the US, DASPOS https://daspos.crc.nd.edu/, is a practical initial exploration to provide data, software and algorithmic preservation for HEP and templates for other disciplines, including the software and policies to understand, trust and reuse the data.

In Europe SCIDIP-ES focuses on Earth Sciences http:// www.scidip-es.eu; also a new project e–ark http://expertpc.org/earkproject/abouteark.htm In co-operation with commercial systems providers, E-ARK will create and pilot a pan-European methodology for electronic document archiving, synthesizing existing national and international best practices, that will keep records and databases authentic and usable over time.

9.2 Earthcube

www.earthcube.org

Establishing interoperability and data and service interface standards is one of the goals of the Earthcube initiative. In particular the building block "GeoWS - Geoscience Web Services aims to "..engage EarthCube cyberinfrastructure in developing, establishing and adopting international standards.."[18]

9.3 OAIS

http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284

NEEScomm is already adopted the OAIS reference model and is benefiting from it.

9.4 OpenFlow Protocol

https://www.opennetworking.org/

The OpenFlow protocol is a standard that supports application level control of networks. OpenFlow enables the development of a Software Defined Networking (SDN) architecture that is dynamic, manageable, cost-effective, and adaptable that decouples the network control and forwarding functions enabling the network control to become directly programmable and the underlying infrastructure to be abstracted.

9.5 Research Data Alliance

https://rd-alliance.org/

The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data. Its work is funded by agencies in multiple continents and its work is accomplished through working groups on particular topics.

10 Funding and Sustainability Models

Several models for cost sharing and acquisition are now being tried out for previously or currently proposed NSF and DOE developments:

10.1 Globus[19]

https://www.globus.org/providers/provider-plans#pricing-plus

Globus depends on the Internet2 Netplus service to for costs related to the data throughput level to be supported. Globus is free for researchers at non-profit institutions to use for file transfers.

The sharing service is available to researchers for a nominal monthly or annual subscription fee. Provider plans are offered that enable resource owners and administrators to deliver enhanced data management capabilities to their users, backed by higher levels of support from the Globus team. Globus Plus subscriptions are just

[18] http://workspace.earthcube.org/geows
[19] Globus Online was recently renamed as Globus.

$7/month (or $70/year) for users at non-profit research and educational institutions, with more specific costs listed at

10.2 HDF Group

http://www.hdfgroup.org/about/business_model.html

The HDF group uses a fundraising model to cover its costs and has funded staff who have some responsibility for this fundraising covering: Comprehensive, long-term support from such organizations as NASA's EOS; Sponsorship through on-going maintenance support; Research and development projects sponsored by organizations to support the mission of The HDF Group; and Targeted projects for organizations from all disciplines that need special services provided. The HDF Group will employ the following approaches for fundraising in support of its mission:

- Governmental grants, and contracts, including cooperative agreements and broad area announcements (BAAs).
- Other public-sector and private-sector grants, contracts, and cooperative agreements.
- Maintenance subscriptions for support and services.
- Referrals from existing grantors, collaborators and/or clients.
- Referrals from HDF technology users.
- Referrals and introductions to potential HDF technology users in scientific, academic, research, and industry sectors, e.g., bio-informatics, hydrology, energy.

10.3 Internet2 Netplus

http://www.internet2.edu/vision-initiatives/initiatives/internet2-netplus/

Internet2 offers a service between the user and the seller to leverage the collaboration across the institutions. http://www.internet2.edu/vision-initiatives/initiatives/internet2-netplus/internet2-net-frequently-asked-questions/

10.4 IPlant Collaboration and Cost Recovery

http://www.iplantcollaborative.org/content/collaboration-policy

IPLant in general does not charge for data or CPU use but has a list of circumstances in which cost recovery will apply. The expectation is that currently this is very little invoked.

10.5 RedCloud

https://www.cac.cornell.edu/redcloud/

The RedCloud charges a subscription based on the CPU hours used which can be extended to include storage needed. There are two types of offerings – the second of which includes the use of MatLab. Cornell faculty and staff complete have a Cornell University account number for charge back and other academic institutions use a credit card or create an invoice.

10.6 Xarxiv

https://confluence.cornell.edu/display/culpublic/arXiv+Sustainability+Initiative

The sustainability initiative implements a business model that is a combination of funding grants and annual fees from member institutions.

11 Software Practices and Research

This section covers potentially beneficial CI research activities as well as software lifecycle and software engineering processes for community driven cyberinfrastructure and collaborative software systems.

11.1 Computing Research

There are many NSF ACI related calls that NEEScomm/IT could benefit from as a user and tester/early adopter. NEES has excellent integration with and attendance at NSF facility meetings such as the Facility Security Meeting; the SI2 PI meetings. A concerted effort to engage the following NSF projects might be of use:

- Partnership Consortia patterned on something like the Open Science Data Cloud which gives access to a common storage and computing facility, open source software developed in collaboration with the partners, and access to an active diverse set of open http://opencloudconsortium.org/ and has membership agreements and licenses fees.
- Collaborating with the NSF funded SI2 projects. These are listed at https://sites.google.com/site/softwarecyberinfrastructure/software/software
- Partner with the National Data Service Consortium http://www.nationaldataservice.org/

11.2 Software Development and Distribution

Code development is a resource intensive, and costly enterprise. Additional tools are being developed to help in community/multi developer communities to improve the initial quality, automated documentation and maintainability of the codes. One key goal of collaborative software systems is to allow many individual or group contributions to occur in parallel and simultaneously and support periodic and/or continuous merging and building of a robust common code base for use by all.

The Fermilab software framework teams (CMS, art) has documented the requirements[20] they work to based on the experience of working in the HEP arena for many years. The goal is to reduce the bottleneck of tight control on upload of new code and validation by one or two individuals by extending automated capabilities, tracking and reporting.

11.2.1 CERN File System (CVMFS)

http://cernvm.cern.ch/portal/filesystem

[20] https://cdcvs.fnal.gov/redmine/attachments/download/18164/requirements.pdf

CVMFS is gaining traction in the field as a way of distributing software to widely distributed computing sites, and providing software repositories where a distributed set of maintainers can upload code for immediate download across the distributed computing facility.

11.2.2 Git

http://git-scm.com/

Git is an open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency. Git is especially powerful at supporting multi-developers by having enhanced "branch and then merge capabilities".

11.2.3 Jenkins

http://jenkins-ci.org/

Jenkins is an open source extendable continuous integration service. Jenkins provides a means for developers to easily include and then exercise regression, unit and integration tests. A recent survey by Fermilab[21] found more than 15 open source packages with the most interesting being Jenkins http://jenkins-ci.org/ http://jenkins-ci.org/content/about-jenkins-ci and BuildBot http://buildbot.net/#/basics

11.3 User Support Tools

11.3.1 Software Carpentry

http://software-carpentry.org/

A community based infrastructure for training of software developers.

11.3.2 XSEDE Extended Support for Communities

The XSEDE project has made effective use of ESCC – or Extended Support for Communities – where experts are embedded into the research community to help make their algorithms and application code work more effectively on the provided computing and storage. (E.g. the LIGO experiment)

11.3.3 UK Software Training Institute

http://www.software.ac.uk/software-sustainability-institute-changing-research-software-attitudes-and-practices

12 How some of the above technologies might be actually useful to and/or adopted by NEES IT

We briefly analyze and comment the utility of the products below based on whether

[21] http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=5348

NEES might actually use the product or service (U) ; whether NEES could benefit from following the pattern of organization, contributions or development models of the project or product (P) ; and whether NEES might usefully partner or engage with the organization or project (E).

Technology	U	P	E	Comments
Amazon Web Services	X			Use. Evaluate and maintain test usage of a small allocation of Amazon cloud resources from NEES toolkit so as to be to use them for peak requests from NEES researchers in the future, and understand both the financial and resource costs of using the services as the purchase model changes.
CIConnect	X		X	CI Connect enables transparent use of local computing and storage resources with remote ones. This is only useful for NEES when/if there is a researcher with this use case, needs and interest to adapt their codes to be successful. The actual adoption and integration should not be more than a couple of months effort given that Batch Submit already submits to XSEDE and OSG resources.
DASPOS			X	DASPOS ends in the next year or so. The outcomes to date have been LHC community specific – an understanding of the needs and how to approach preserving both the software as well as the data. Of interest is to understand the next steps and see whether any collaboration with NEES would be useful. SCI-DEP is a European data preservation project.
DOCKER	X			This is a no brainer based on the increasing level of adoption. Evaluate utility by interfacing to NEEShub tools for submitting work to virtualized resources.
DOE SciDac Institutes			X	The DOE SciDAC institutes develop applied mathematics and data analytics to benefit DOE science communities. An Engagement by NEES with selected institutes to discuss some of the specific analysis scenarios of the NEES researchers would be a good step in understanding whether there are potential benefits from increased engagement.
Dropbox	X			Can be used without specific NEES support and integrated into the NEES IT toolkit. If used for data greater than a few 100s Gigabytes will require licences; Interfacing to PEN will involve development.
Earthcube Building	X		X	Several members of NEES are already engaged with

Technology	U	P	E	Comments
Blocks				Earthcube, there have been ongoing discussions of NEES PAC and Executive Council about participation. The overhead of engagement is high in terms of meetings, discussion, adopting the terminologies and approach. It still remains to be seen how generally useful or usable any deliverables from the Building Blocks projects are. We have had ongoing discussions about collaborating on cyberinfrastructure and software tools in the Cyberinfrastructure Committee bewtee IRIS and NEES without any tangible productive outcomes that I am aware of. Only recommendation can be to take the existing NEES software wishes and strategic goals and map them onto the building block project goals and look for overlaps; and or have a focused technical workshop that involves CI knowledgeable NEES researchers and Earchcube researchers.
Globus	X	X		
Google Drive	X			Another product that can be used without specific NEES support and integrated into the NEES IT toolkit. If used for data greater than a few 100s Gigabytes will require licences; Interfacing to PEN will involve development. Dropbox has longevity in the field.
HYDRA	X		X	The HYDRA model of development and community looks to have had some major successes. A NEES research example, using the HYDRA tool kit for data management and storage, might be of great benefit to our community.
InCommon	X		X	The InCommon Federation provides certificates and trust services in support of Single Signon.
IPLANT collaborative		X	X	NEES and IPLANT have similar models in serving their specific scientific/research domains. They adopt common community underpinnings such as Globus, Condor and DOIs. They implement services and products based on the prioritized wishes of the communities they serve; and work with the researchers to support their applications and infrastructure needs in whatever state of maturity they currently are. Several members of NEES are already engaged with IPLANT. The Agave toolkit can be compared to Hub0 and its “plug-ins” If software components can be pugged into both engagement

Technology	U	P	E	Comments
				between the projects could be quite productive for the longer term.
LHCONE		X	X	LHCONE as adopted an "overlay" approach to network provisioning to give the application community a transparent interface and control plane over a federated, diverse set of network providers for the distribution and movement of large data sets. Additionally, LCONE provides a working example of a domain science community coming together "as a collaboration without direct funding" to engage a technology provider (ESNET) to meet their needs.
National Data Service			X	The recommendation is "Engage" here because of the immaturity of the Consortium, NEES should develop a list of questions to be asked in the initial approach, including: better understanding of the model proposed; partnership aspects to ensure the best ROI for the efforts to be applied; and the planned longer term operating goals e.g. how does the member maintain independence to choose and implement their own solutions; etc.
Neon		X		The comment is similar to that for IPLANT with the exception of whether there is utility in exchange of datasets and information between the projects.
NSF Scientific Gateways			X	The next phase of the NSF Scientific Gateways is likely to continue the current practice of a discussion and exchange of information forum and providing short term injection of expertise to projects on request, as well as potentially a place for portals to be hardened and supported for the scientific community. It would be useful for Hub0 and NEESHub to participate as one of the software solutions supported.
Open Science Data Cloud		X		Learning from the pattern of organization and approach in terms of gaining the interest of commercial investors, forming a not-for-profit and requiring membership fees for use of the software, hardware and consulting, that seems the most useful thing that can be learned from this project. (gven the existing NEEShub, storage available at Purdue and technologies in use)
Open Source Data Turbine	X	X	X	NEES already has an collaboration here through Shirley Dyke who will know more than the writer of

Technology	U	P	E	Comments
				this report what the useful options are.
ORCID	X			This is a no-brainer. In parallel with DOIs becoming universal Document Identifiers, ORCID is growing to become universal "Human identifiers" for research and science. Join!
RedCloud		X		The outcomes of this project have been to demonstrate a working model for providing access to both single-node and multi-node Matlab applications, based on a centralized model of Matlab licences. NEES has already made one step towards this by providing local matlab resources at Purdue. Extending this to involve researchers using local Matlab licences at NEES sites and/or other universities and establishing a releationship with the vendor would provide NEEScom ROI in terms of the number of researchers that can be supported.
Research Data Alliance			X	This has become the recognized international forum for talking about data management and provenance attributes, standards, semantics and descriptions. It does not itself provide software or other artifacts but documented standards, best practices and recommendations. It is important that stay aware of its activities and outputs; It is becoming increasingly a useful forum for the exchange of ideas and discussion – with relatively low investment of resources.
SCIDIP			X	SCiDIP is a European project scheduled to end in 2014 with expected continuation by the Members. It would be useful for NEES to continue to engage
Seedme	X		X	Seedme is based on a python and yt toolkit. A technical evaluation of the platform by NEES IT would be useful to understand potential benefit from collaboration and sharing of software.
Software Carpentry	X		X	Software carpentry training and education classes as well as organization has increasing name recognition. Many projects are finding value in associating with the name and the model. Software carpentry provides the common underlying software and CI education, the domain needs to extend these to their own use cases and technologies. It is "merely" a decision of approach and investment directions for the NEES training and education arm. This inevitably depends on the NEES individuals involved.

13 Appendix: What is Cyberinfrastructure?[22]

There are many definitions for the term cyberinfrastructure. Based on the features mentioned in the 'NSF Cyberinfrastructure Vision for 21st Century Discover', cyberinfrastructure includes:

- Computing systems
- Data
- Information resources
- Digitally enabled sensors and Instruments
- Virtual organizations
- Interoperable suite of software services and tools

[22] http://nees.org/about/NEEScomm/cyberinfrastructure